POETIC: Interactive Solutions to Alleviate the Reversal Error in Student-Professor Type Problems

Sung-Hee Kim[a], Daniel Phang[b], Tuyin An[c], Ji Soo Yi[a], Rachael Kenney[c,d], and Nelson A. Uhan[e]

[a]*School of Industrial Engineering, Purdue University, West Lafayette, IN 47906*
[b]*P.C. Rossin College of Engineering, Lehigh University, Bethlehem, PA 18015*
[c]*Department of Curriculum & Instruction, Purdue University, West Lafayette, IN 47907*
[d]*Department of Mathematics, Purdue University, West Lafayette, IN 47907*
[e]*Mathematics Department, United States Naval Academy, Annapolis, MD 21402*

Abstract

The *reversal error*—reversing the relationship between two variables in a mathematical word problem—is a long-standing issue in mathematics education, despite its apparent simplicity. In this paper, we describe and study POETIC, an interactive web-based environment we developed to teach users to avoid the reversal error. POETIC uses two types of novel interactive visualization, called the Test-Case and Room-Metaphor approaches. To verify the effectiveness of these approaches, we conducted crowdsourcing-based comparison studies with 200 participants and found that both approaches significantly decreased the frequency of reversal errors for certain types of word problems. Our results show that interactive visualization of equations can reduce the occurrence of the reversal error.

Keywords: mathematics education, reversal error, interactive visualization, test-case, room-metaphor

1. Introduction

In 1979, Kaput and Clement wrote a letter to the editor of *Journal of Children's Mathematics Behavior* introducing the "student-professor problem" to describe the difficulties that students face while solving mathematics word problems. The problem was originally worded as follows:

> "Write an equation using the variables S and P to represent the following statement: 'There are six times as many students as professors at this university.' Use S for the number of students and P for the number of professors." (Kaput and Clement, 1979, p. 208).

In spite of its simplicity, this problem and its variations (hereby referred to as *SP-type problems*) have caused difficulty for students at all education levels. Several studies have reported that roughly 40% of college students fail to solve these problems correctly (Clement, 1982; Fisher, 1988; Lochhead, 1980; Philipp, 1992; Sims-Knight and Kaput, 1983; Weinberg, 2009; Wollman, 1983). We have also encountered this phenomenon while investigating common errors that students make while constructing mathematical optimization models in a junior-level undergraduate engineering course (Kenney et al., 2011).

Previous research has shown that the most common error students make with *SP*-type problems is the *reversal error* (Sims-Knight and Kaput, 1983), in which students reverse the order of variables; for example, for the above word problem, writing $P = 6S$ instead of the correct answer, $S = 6P$. This error has been observed consistently among many research participants. Over the past 30 years, many studies have tried to determine why students consistently commit the reversal error and how to correct this error (see Section 2 for more details). However, to

date, there have been mostly mixed results from various studies: no method has been consistently successful.

Drawing on the literature in mathematical education, we believe that this reversal error is not an issue of mere carelessness, but it is a more deeply rooted problem of how students comprehend the problem and interact with mathematical notation; in other words, there may be a cognitive incompatibility between a problem solver's internal representation of the problem (or a mental model according to Liu and Stasko (2010)) and the mathematical representation (Weinberg, 2009).

Thus, the goal of our research is to develop and evaluate a novel web-based interactive educational tool that will help reduce the occurrence of the reversal error in *SP*-type problems by helping students better relate their mental model of the problem to the mathematical notation they use to represent it. To achieve this goal, we first thoroughly reviewed the existing literature and organized findings to inform our study design. We conducted pilot interviews with ten students to better understand why they made reversal errors and prototyped multiple design alternatives. Then, we developed an interactive visualization tool, called "POETIC," with two novel interaction techniques—the *Test-Case* and the *Room-Metaphor* approaches. We tested their effectiveness through crowdsourcing-based studies.

2. Background

2.1. Potential Causes for the Reversal Error

Several studies have investigated reasons for students' tendency to make the reversal error. They have found that students often focus on the text description and lack mathematical understanding of the equations.

2.1.1. Direct Translation and Static Comparison Strategies

Many researchers have attributed student difficulty to a tendency for a "direct-translation" approach (Clement, 1982; Fisher et al., 2010; Hegarty et al., 1995; Wollman, 1983). This strategy is used when students attempt to make the sequence of algebraic symbols match the sequence of objects in a word problem. For example, when the student-professor problem is written as "There are six times as many students as professors," students make the reversal error by translating directly to 6 times S equals P. Fisher et al. (2010) suggests teaching students that the "standard" multiplicative format for equations (e.g., $ax = y$) could be a contributing factor that encourages direct translation.

However, even when phrasing questions so that a direct translation would produce a correct response (e.g. "The number of students is six times the number of professors"), students still tend to make the reversal error (MacGregor and Stacey, 1993; Stacey and McGregor, 1993). Using a "static comparison" strategy, "$6S$" and "P" are treated as the objects "six students" and "one professor" respectively (instead of as the *number* of students and professors). The equal sign is also seen as representing correspondence (6 students equals 1 professor) rather than equality (Cohen and Kanim, 2005; Palm, 2008). Researchers have conjectured that students who commit the reversal error may have a deep cognitive bias that results in these static images being formed (Kaput and Clement, 1979; MacGregor and Stacey, 1993; Stacey and McGregor, 1993).

2.1.2. Impacts of Problem Description

Variations in problem (or task) descriptions have also been reported to significantly impact students' performances on *SP*-type problems. For example, the wording sequence can impact performance. Students often make more reversal

errors with problems containing phrases like "1 cow for every 6 pigs" in which the sequence of words does not match the sequence of variables in the correct mathematical expression (Clement, 1982; Cohen and Kanim, 2005; MacGregor and Stacey, 1993; Philipp, 1992; Stacey and McGregor, 1993). Sims-Knight and Kaput (1983) also found that problems that contained a context of imageable words, such as number of students, posed more difficulty than problems with non-imageable words, such as height of Mount Everest. Philipp (1992) observed that the success rate for a "familiar" problem, in his example a problem that asked for an expression that relates the values of stacks of pennies and dimes, was only 11%. This is much lower than the success rate of the original student-professor problem, which was 34-63% (Clement, 1982). Additional difficulties arise when the use of coefficients other than "1" for one of the variables cause an issue of divisibility. For example, Clement observed that students had significantly more trouble with *SP*-type problems involving a non-trivial ratio (e.g., "four cheesecakes for every five strudels"), with a success rate of 27%.

We have focused on these different types of problem descriptions, i.e., Wording Sequence (Type 1), Imageability (Type 2), Familiarity (Type 3), and Divisibility (Type 4), to inform the development of the questions used in our experiment.

2.2. Prior Recommendations from the Literature

In the past 30 years, several solutions to alleviate the reversal error have been proposed and evaluated. While no single fix has been identified, these reports serve as a foundation for our continued efforts to help students recognize and overcome reversal errors.

2.2.1. The Use of Visual Representations

Visual perception is one of the basic human sensory sources of mental models which encodes the external world into internal representation. In this sense, visualization can be a powerful way to understand abstract concepts by representing them in an explicit way. While reviewing human-computer interaction and other relevant domains, we have identified some interesting ideas related to visually representing mathematical equations (e.g., Scrubbing Calculator [1], Mortensen Math blocks [2]), but none of them directly handle the reversal error and we have found no corresponding empirical studies which might help us understand the potential of such approaches.

Instead, we identified an interesting thread of research using visualization techniques to accompany algorithms in computer science (Lawrence, 1993; Mulholland, 1998; Stasko et al., 1993; Tung et al., 2001), which tackles a similar problem. Hundhausen et al. (2002), for example, investigated how visualizations were used and in what context they were found to be useful in teaching algorithms. After analyzing 24 experimental studies using various visualization techniques to teach algorithms, the authors found that visualizations that lead to successful educational outcomes share two underlying theories: The first one is *epistemic fidelity theory* (Hundhausen and Douglas, 1999), which emphasizes that there is a right mental model (often an experts' mental model) for specific reasoning and action; the better the graphical representation or visualization fits the expert's mental model, the more efficient the transfer is to the viewer who decodes the internalized target knowledge. The other theory is *cognitive constructivism*, which emphasizes that individuals

[1] `http://worrydream.com/ScrubbingCalculator/`

[2] `http://www.mortensenmathdirect.com/`

construct their own knowledge through experiences (Resnick, 1989). While being engaged, individuals construct new understandings by interpreting new findings and combining these with their own existing knowledge. This theory emphasizes that active learning is important for changing one's mental model. No matter how high the level of the epistemic fidelity is, passively viewing visualizations may not be sufficient for building conceptual understanding. These two theories were inspirational to our design procedure.

Some studies in the mathematics education literature have investigated the role that visual representations can have on students' modeling of word problems. In an early study, Sims-Knight and Kaput (1983) used pictorial representations when administering *SP*-type problems. Interestingly, the error rate was much worse: 60.4% of undergraduate students got the problem wrong with the visual aid, whereas 37% got the problem wrong without the visual aid. More recent studies, however, tend to disagree with this early finding. Studies of experienced mathematicians shows that they often make use of diagrams to aid in tasks of mathematical analysis (Stylianou, 2002; Waisel et al., 2008). Yazdani (2008) found that asking students to draw a picture, figure, table or graph greatly improved students ability to grasp problems on which reversal errors had previously been made. Visualizations may help students detect errors and see limitations in previously used strategies such as direct translation. In this study, we focus on a visualization of the mathematical expressions themselves to enhance the understanding of the quantitative relationships in an effort to further enhance students' abilities to solve word problems without the reversal error.

2.2.2. Behavioral Patterns of Successful Problem Solvers

Early work by Clement (1982) identified two distinct behaviors or patterns seen in students who can successfully solve *SP*-type problems. An *operative approach* is used when a student completes a hypothetical operation that produces an equivalence relation. As equations written in a standard multiplicative format do not describe the situation at hand in a direct manner, students have to understand the equivalence relationship in mathematical terms. Some operative approach mental solutions include $S/6 = P$ or $S/P = 6$ (Palm, 2008). Fisher et al. (2010) found that requiring students to write a non-standard relationship by developing such equivalence relationships (e.g., "If I divide the number of students by 6, I get the number of professors") decreased the appearance of the reversal error significantly. The second pattern of successful word problem solving is what Clement called the *substitution pattern*. Here, students substitute some numbers which fit the described situation into the variables of the constructed equation, and then engage in the operative approach described above. Students using the substitution pattern appear to understand the equations in terms of their relations with concrete numbers, and if an equation with substituted numbers (e.g., $6S = P \Rightarrow 6 \times 6 = 1$) turns out to be false, they reconsider the validity of the equation. Behaviors that can lead to correct responses such as the two described here could be the key in helping students overcome reversal errors.

For this study, we suggest that if a tool could promote those behaviors, the occurrence of the reversal error could be reduced. In this paper, we investigate students' success when encouraged to engage in these patterns through our Room-Metaphor and Test-Case approaches (see Section 3.2).

2.3. Next-Day Phenomenon

As we incorporate the ideas above into our study we wish to be mindful of what, in the context of cognitive and learning issues, Tzur and Simon (2004) refer to as the "next-day phenomenon." For example, consider young students who successfully determine if 1/6 or 1/8 is the larger fraction by partitioning paper strips, but when asked the next day to determine the relation between 1/7 and 1/3 without the paper strips, are unsuccessful. The authors posit that this is not an issue of forgetfulness, but that for these students, knowledge is only available within the context of the activity in which it was created. In this situation, students may be able to re-engage their previous understanding if prompted to connect back to the activity of using the paper strips, but they have not yet reached a stage of being able to anticipate a connection between the paper manipulation and the mental act of comparing fractions. When students have a fully developed a proper mental model that was initially developed through some activity (in our study this will be in the context of the web-based tool), then they may be able to independently "anticipate" the knowledge without being limited to a certain activity.

We attend to the next-day phenomenon in our work to ensure that learning is not restricted within the context of the activity in which we engage the students. To do this, we need to test the effectiveness of our tool a few days after the learners' initial interaction with it to see if transfer has occurred outside the original activity.

3. Design

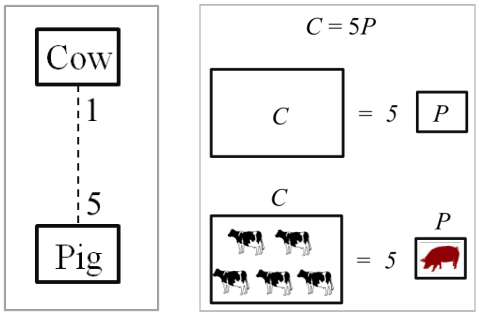
For this study, we adopted a user-centered design approach (for a review, see Vredenburg et al., 2002) to design solutions for *SP*-type problems. In pilot studies, we observed how students solved problems to try to start to understand

their difficulties. Several initial ideas, informed by different ideas used in previous literature, were tested using paper prototypes in interview studies with student participants. These interviews were informal and the interview questions and paper prototypes evolved over each interview session. However, the interview results helped to confirm findings from previous literature and also design the new interactive tool.

3.1. Interviews

We conducted interviews with ten undergraduate and graduate engineering students (four female) who were provided with various types of *SP*-type problems with similar levels of difficulty and different contexts. All problems were designed to include statements that could lead students to potentially make the reversal error. Students explained their thought processes as they solved. We found that students who answered the questions correctly used various strategies, such as a substitution pattern, a translation of the equation back into words, and a method of setting up the given information using ratios. In contrast, all the students who committed the reversal error considered the variable as an object label.

In order to help students solve the problem, we promoted a substitution pattern by suggesting students write test cases (e.g., $(S = 6, P = 1)$, $(S = 12, P = 2)$, etc., for the student-professor problem) before constructing an equation. In general, students were comfortable with writing test cases after reading the word description, and all students generated correct test cases. However, even with correct test cases, most still made mistakes in writing the equation. The test cases did help six students realize that there was something wrong and eventually helped them to correct the equation. However, other students did not benefit from the test cases. For example, upon realizing that the test cases and the written equation were not compatible, one



(a) (b)

Figure 1: Visual representations used during the interviews for the problem "For every five pigs, he raises one cow." (a) A node-link diagram to help students organize the word description. (b) A visual representation of an equation when the equation contains the reversal error.

student's first attempt was to modify his test cases instead of rewriting the equation. This indicates that providing simple feedback showing that something is wrong may not sufficiently help students understand *why* it is wrong.

Therefore, our next approach was to try various visual representations of problems and mathematical notation. We showed participants a node-link diagram (see Figure 1(a)) to represent two objects and the relationships between the two (i.e., one cow for every five pigs). However, Figure 1(a) actually made students more prone to making the reversal error. This reaffirmed a need to develop visualizations that focus on the mathematical expressions to help break the notion of a variable corresponding to an object (instead of number of objects). We next created Figure 1(b), an example showing what an equation with the reversal error actually means by showing one pig and five cows. This visualization seemed to be better understood by participants in the interviews, which encouraged us to build on this idea for our tool design.

3.2. Design Ideas

After several iterations of design changes, we came up with two novel approaches that warranted further investigation: (i) the *Test-Case approach*, and (ii) the *Room-Metaphor approach*. These two features were implemented and combined together to provide proper feedback to students as they interact with *SP*-type problems.

3.2.1. Test-Case Approach

The interview students' successes with constructing correct test cases suggests that the mental model of the problem situation is not necessarily an issue for the students. However, they were not always able to translate this mental model to a correct mathematical expression. To harness students' comfort with test cases, we designed a feature of our tool to trigger students' awareness when a discrepancy arises between test cases and constructed mathematical expressions. The inclusion of this feature is supported by Clement (1982) who found higher students success with *SP*-type problems when using a substitution strategy. However, we failed to find in the literature any effective mechanism that encourages students to use such a strategy except through direct interaction with an instructor. Our interface prompts students to come up with test cases and provides instant feedback when the test case and equation are not compatible. This approach is similar to test-driven development or unit-test approaches used in software engineering (Beck, 2003) as test cases that students develop could drive equation construction. This idea is also in line with cognitive constructivism because students are led to interact with both an equation and test cases.

3.2.2. Room-Metaphor Approach

To address situations where test cases do not rectify initial mental models of a problem, we developed a second, more visual approach to aid students' understanding of the underlying mathematical mechanism. Inspired by epistemic fidelity theory, we created a visual representation in our tool that alludes to the notion of variable as a "container" of a number of objects rather than as an object itself. We call this the "room metaphor," where a variable is visually represented as a room that holds some number of boxes. As suggested by previous studies, students who struggle with *SP*-type problems often apply the static comparison approach where variable letters treated as objects (e.g. S= students) instead of quantities. Our hypothesis is that the room metaphor can help break this static model and, in turn, help students to better understand the equation structure. This is also in line with the epistemic fidelity theory since the room metaphor could be closer to the ideal mental model that students should have.

3.3. Prototypes

Based on the two approaches, we developed prototypes of our tool. The main idea was to harness the input of students where they do not make mistakes (i.e., Test-Case approach) and try to provide external metaphors to the students to understand the equivalent relationship of the equation (i.e., Room-metaphor approach). First, to implement the Room-Metaphor approach, we visually represented a variable as a block of arbitrary color encapsulating the variable letter as Figure 2 shows.

When there is a clear relationship between variables, as in the student-professor problem and other *SP*-type problems, the size of the visual blocks or "rooms" can be calculated by the coefficients in front of these variables. For example, for the equation $S = 6P$ in the student-professor problem, the value for the variable S is

Figure 2: A variable, P, is represented as a visual block.

(a) The visual block for S has a size that is *six times bigger* than the visual block for P, to be consistent with the equation.

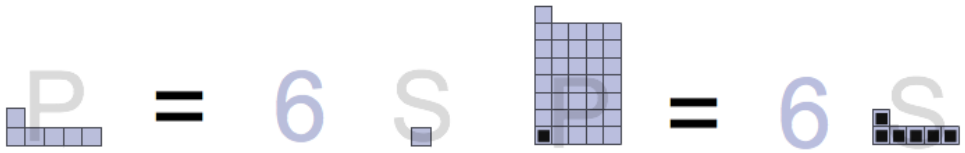
(b) The number of black boxes are drawn based on the sample values and fill the visual blocks with no empty spaces. (Sample values: S = 6 and P = 1)

Figure 3: Example of (a) Room-metaphor and (b) Test-Case feedback with correct equation and correct sample values.

clearly *six times* the value for the variable P. Accordingly, as shown in Figure 3(a), the room corresponding to S is six times bigger than the room corresponding to P. Note that because this is an equation, the total "room" size on each side of the equation must equate. In this case, the size of the S visual block is equal to the size of the P visual block, multiplied by its coefficient of six.

In addition, when test cases are specified for these variables as discussed in the Test-Case approach, black "boxes" will appear in these visual blue "rooms" to show that the variable has taken on a value. As shown in Figure 3(b), if the sample values correctly satisfy the equation, the black boxes will fill the blocks perfectly with no empty spaces. For example, if there is no mismatch, such as when $S = 6$ and $P = 1$ for the equation $S = 6P$, there are no empty spaces.

For a wrong equation, such as $P = 6S$, the visualization will appear as in



(a) Visual representation for the incorrect equation $P = 6S$ without any test-case values associated.

(b) The visual block size is calculated with the sample values where the empty spaces indicate that the equation and the sample values do not match. (Sample values: S = 6 and P = 1)

Figure 4: Example of (a) Room-metaphor and (b) Test-Case feedback with incorrect equation and correct sample values.

Figure 4(a), where the areas of the empty rooms do not indicate that there is a mistake. However, when combined with a test case (e.g., $S = 1$ and $P = 6$), the feedback is as shown in Figure 4(b). The visual block size over variable P is determined by the equation and the sample values which becomes 36. By drawing the black boxes from the sample values, the empty space works as a visual cue to indicate that there is a mismatch between the values supplied and the equation, i.e., that the test case and the equation do not match.

3.4. Implementation

To implement our combined Test-Case and Room-Metaphor approach, we used the Adobe Flash platform (ActionScript 3.0), which makes the system easily web-accessible. The tool, called "Purdue Optimization Modeling education Tool – Interactive equation Component (POETIC)," is part of a larger system and project, "Purdue Optimization modeling Education Tool (POET)," aimed at helping students to correctly solve mathematical optimization modeling problems. The initial

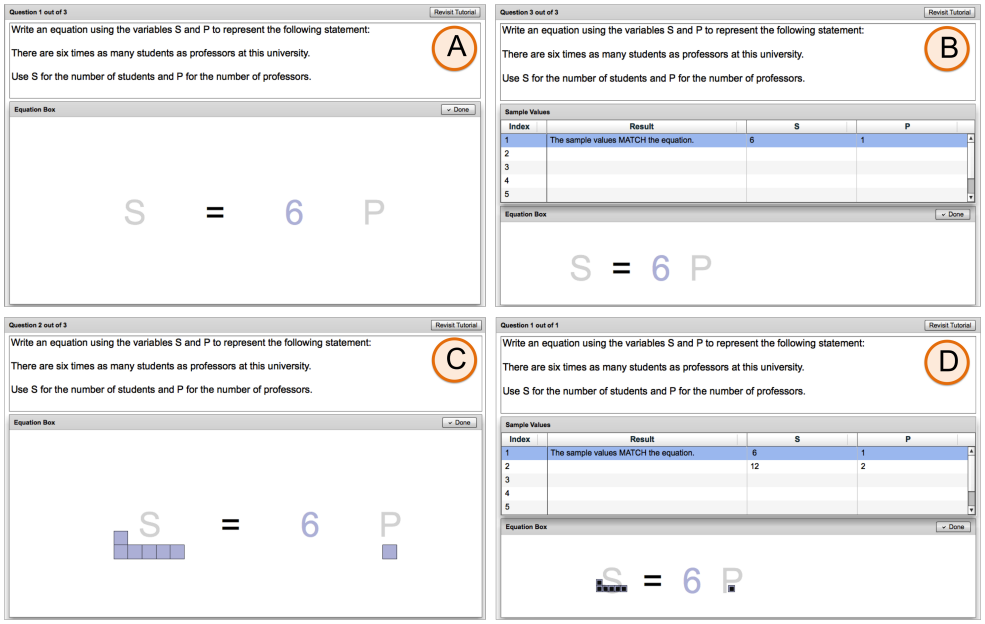


Figure 5: Screenshots showing the four different variations of the POETIC interfaces: (A) The Baseline condition, (B) The Test-Case Condition, (C) The Room-Metaphor Condition; and (D) The Both Condition.

version of the tool, POETIC v0.1, works as an equation editor that provides interactive visualization feedback using our two approaches.

For the experiment, four variations of the POETIC interface were created as shown in Figure 5. Depending on the experimental condition, the POETIC interface contains or does not contain an editable table called *Sample Values* (used for test-cases) on top of an *Equation Box* for equation editing. We allow only one equation to be entered at a time, because all of the *SP*-type problems we are concerned with involve a simple one-equation, two-variable relationship. For the Equation Box, the sizes of blue blocks are automatically adjusted depending on the equation entered by the user as shown in Figure 5C and Figure 5D. When the Sample Values table is present, the tool also provides feedback when the user

hovers the mouse on one row of the Sample Values table; in particular, POETIC reports "The sample values (MIS)MATCH the equation." as shown in Figure 5B and Figure 5D. Feedback using the black boxes, as shown in Figure 5D, only appears when both the Test-Case and Room-Metaphor conditions are activated for the user.

4. Hypotheses

As suggested by the four interface designs shown in Figure 5, the overarching research question of this study is whether the two approaches (i.e., the Test-Case and Room-Metaphor approaches) are effective on reducing reversal errors. Specifically, we had the following five hypotheses.

H1 The Test-Case approach reduces the occurrence of the reversal error.

H2 The Room-Metaphor approach reduces the occurrence of the reversal error.

H3 There is an interaction effect between the Test-Case approach and the Room-Metaphor approach.

H4 The Test-Case approach is subject to the next-day phenomenon.

H5 The Room-Metaphor approach is not subject to the next-day phenomenon.

In summary, we hypothesized that both approaches would be effective in reducing the occurrence of the reversal error (H1 and H2), but the Test-Case approach would be subject to the next-day phenomenon while the Room-Metaphor approach would not (H4 and H5). We also hypothesized that the two approaches would create a synergy when both of them are used together (H3). The experiment

was divided into two phases. User Study 1 (Section 5) tests the effectiveness of the two approaches following with User Study 2 (Section 6) to test the longitudinal effects to see if the learning during User Study 1 has transferred.

5. User Study 1

User Study 1 was designed to test H1, H2, and H3. Since a crowdsourcing-based user study allows us to collect large amounts of data in an economic manner (Kittur et al., 2008) from a diverse population, we used Amazon Mechanical Turk to perform our experiment.

5.1. Methods

5.1.1. Participants

We recruited a total of 200 participants (79 females) for three days, and each participant was randomly assigned to one of four conditions (50 participants per each condition). The participants' ages ranged from 18 to 65, and the average age was 28.6. The incentive for completing the whole experiment was $0.30 (which is typical for a task on the Mechanical Turk platform), and the average time required to finish the whole experiment was 28 minutes. Education level was relatively high: the experiment included participants with a 4-year college degree (28%), a 2-year college degree (28%), and a master's degree (19%). The participants had backgrounds in computer and information science (24%), engineering (17%), science and math (13%), agriculture and related science (10%), and business (8%).

5.1.2. Procedures

In order to evaluate our two approaches, we randomly assigned our participants into one of four conditions:

- baseline with no feedback (*Baseline*, Figure 5A);
- only Test-Case feedback (*Test-Case*, Figure 5B);
- only Room-Metaphor feedback (*Room-Metaphor*, Figure 5C);
- both Test-Case and Room-Metaphor feedback (*Both*, Figure 5D).

Each participant solved eight problems. To measure the effectiveness of each interface, the first four problems given to all participants (regardless of their condition) were in the Baseline condition (Figure 5A). We call these first four problems the "Baseline Quiz." After a participant completed the Baseline Quiz, he or she was given another set of four problems in his or her randomly assigned condition, which we call "Quiz 1." Instructions for each interface were given right before the Baseline Quiz and Quiz 1, and participants could revisit the instructions at any time during the task by pressing the "Revisit Tutorial" button on the top right corner of the screen.

5.1.3. Task

The task was to read the word problems and type a response (a mathematical equation) in the Equation Box. In the Test-Case and Both conditions, if a participant did not use the Sample Values table (i.e., did not enter any test cases) and pressed the "Done" button to move on to the next problem, the participant was shown a warning message asking her to use the Sample Value table at least once.

Because prior literature showed that subtle variations of wording impact students' performances, we intentionally designed four different types of problems to cover the variations of *SP*-type problems. Both the Baseline Quiz and Quiz 1

Table 1: Problem descriptions

Problem Type	Quiz	Problems
Type 1	Baseline	There are six times as many students as professors at this university.
(Baseline)	Quiz 1	A country sells four times as much wheat as corn.
	Quiz 2	There are five times as many boys as girls in a classroom.
Type 2	Baseline	Mount Everest in Tibet is three times higher than the Alps in Switzerland.
(Imageability)	Quiz 1	The Niger is three times as long as the Rhine.
	Quiz 2	The Eiffel Tower is six times higher than the Leaning Tower of Pisa.
Type 3	Baseline	The value of the pile of pennies is as much as that of the value of the pile of dimes.
(Familiarity)	Quiz 1	The Simplex Company manufactures 1 tabletop for every 4 legs.
	Quiz 2	The White Company produces winter gloves that have a palm piece and five finger pieces.
Type 4	Baseline	A farmer has found that over the years, for every eight pigs he raises, he raises five cows.
(Divisibility)	Quiz 1	At Mindy's restaurant, for every four people who ordered cheesecake, five people ordered strudel.
	Quiz 2	At Abby's restaurant, for every six people who ordered chicken salad, four people ordered pasta.

had of for each of the four types of *SP*-type problems discussed in the Background section (note that the exact same problem was not given in the two quizzes, just problems of the same type). The order in which the problem types were presented was different for the Baseline Quiz and Quiz 1. The problem statements used for each type of *SP*-type problem are shown in Table 1.

One limitation in our study design is that we did not include an additional baseline question at the end of User Study 1 to test for knowledge transfer. The Quiz 2 questions shown in Table 1 were used in User Study 2 (see Section 6) to address transfer and the next-day phenomenon; however, adding a transfer question in this first collection of data may have provided additional useful information for our analysis.

5.1.4. Measures

The web-based experimental system recorded all of the text inputed by participants in the Equation Box along with time stamps. The final equation entered was used to measure their correctness (i.e., correct = 1 and incorrect = 0). Because students have individual differences on their prior knowledge, we used performance improvement from the Baseline Quiz to Quiz 1 instead of using the correctness of Quiz 1. Since the Baseline Quiz and Quiz 1 have the same four types of problems, we codified improvement as follows for each problem type: 1 if we see improvement (Baseline Quiz incorrect and Quiz 1 correct) or if the participant is already capable of solving the problem (Both Baseline Quiz and Quiz 1 correct), 0 if we see no improvement (Both Baseline Quiz and Quiz 1 incorrect), and -1 if we see a negative effect (Baseline Quiz correct and Quiz 1 incorrect). Thus, each participant in User Study 1 was assigned four improvement scores, one for each of the four problem types. A post-experiment survey was also administered over the web

to collect basic demographic information (e.g., age, gender, and major) and any comments regarding the experiment.

5.2. Results

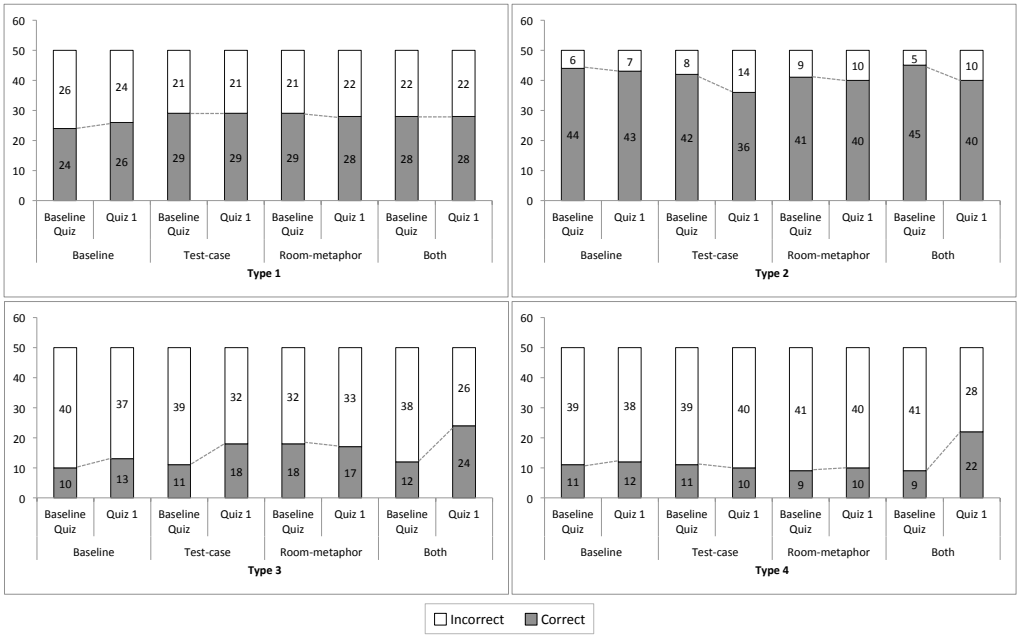


Figure 6: Number of correct and incorrect responses for each problem type and condition in User Study 1. The responses are divided into four problem types and each problem types is shown by condition with Baseline Quiz and Quiz 1. The bar height is the total number of participants for each condition (i.e., 50 participants) and the grey shaded area shows the number of correct responses. The dotted lines between the results are added to emphasize the relative changes between Baseline Quiz and Quiz 1.

Sims-Knight and Kaput (1983) reported that the most common error (roughly over 2/3) that appeared in their study on *SP*-type problems was the reversal error. Our results showed a more skewed pattern: the most common incorrect responses

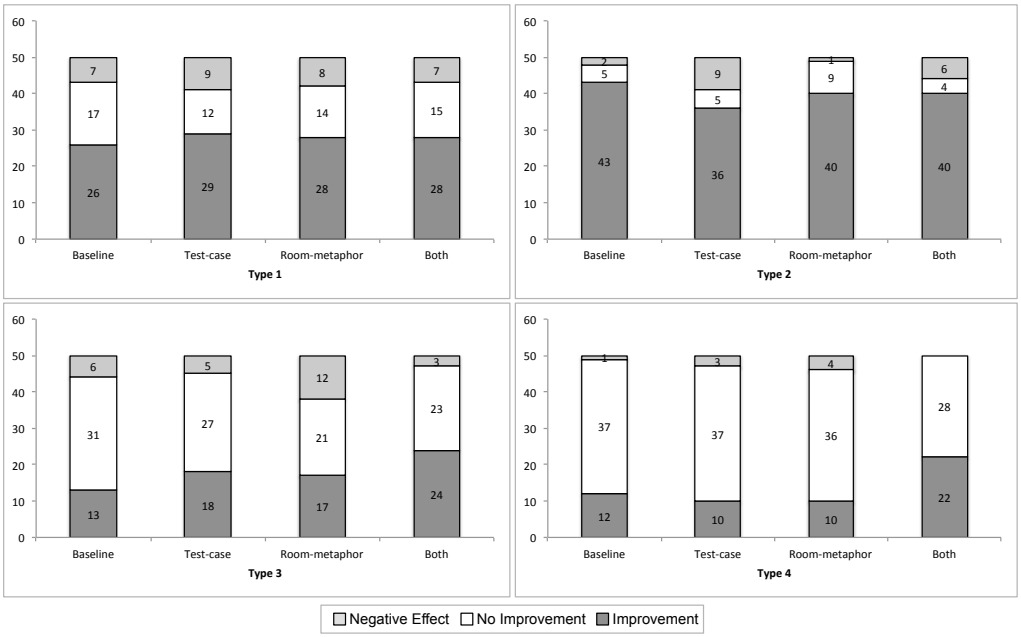


Figure 7: Relative changes (i.e., Negative Effect, No Improvement, and Improvement) in responses for each problem type and condition in User Study 1. The responses are divided into four problem types. The bar height is the total number of participants for each condition (i.e., 50 participants).

were caused by reversal errors (91.05%); the second most common error (4.87%) involved additive expressions without an equal sign (e.g., $6S + P$).

The number of correct answers are shown in Figure 6. As results differed between the problem types (Types 1, 2, 3, and 4), the results for each problem type are plotted on a separate panel. In each panel, the results were divided into four different conditions (Baseline, Test-Case, Room-Metaphor, and Both), and the results of Baseline Quiz and Quiz 1 are shown next to each other. We also added dotted lines between the two results to emphasize the improvement or degradation (i.e., relative change) of participants' performances.

Since the relative change between Baseline Quiz and Quiz 1 was the primary

interest of this study, further analysis was mainly based on the improvement measures (-1: Negative Effect, 0: No Improvement, and 1: Improvement), discussed in Section 5.1.4 (see Figure 7). The improvement measure was an ordered but categorical measure, so a parametric statistical test could not be used. Instead, an ordered logistic regression that handles ordinal multilevel responses using a proportional odds model (Stokes et al., 2000) was employed to determine the effect of the conditions among different problem types.

First, considering all of the problem types together, we found that both the Test-Case and Room-Metaphor interfaces had significant main effects (Wald Chi-Square = 4.3907, $p = 0.0361$ and Wald Chi-Square = 5.3643, $p = 0.0206$, respectively) with a marginal interaction effect (Wald Chi-Square = 3.6247, $p = 0.0569$).

Second, since the results differed between the problem types, as can be seen in Figure 6, the resulting data were divided by their problem types and separately analyzed. For Type 1 and Type 2, no conditions were found to have a significant effect ($p > 0.05$ for any cases). For Type 3, only the Test-Case feedback had a significant effect (Wald Chi-Square = 5.3825, $p = 0.0203$). For Type 4, both the Test-Case and Room-Metaphor feedback had significant main effects (Wald Chi-Square = 8.8555, $p = 0.0029$ and Wald Chi-Square = 7.8661, $p = 0.0050$, respectively) with a significant interaction effect (Wald Chi-Square = 6.8541, $p = 0.0088$).

In addition, the total time spent solving each problem was also measured, and a repeated measures Analysis of Variance (ANOVA) test was employed to understand which factors had impact on the solving time.

First, to see the impact of the two feedback approaches, the time spent on different conditions in Quiz 1 was analyzed. Although three participants could

have been considered as outliers as their times were more than 5 standard deviations away from the mean, we did not remove them as those outliers did not change the results of the analysis. The average time spent for each condition was (from lowest to highest): Baseline (50.0 sec), Room-Metaphor (64.6 sec), Test-Case (131.5 sec), and Both (171.1 sec). Both the Test-Case ($F(1, 597) = 55.27$, $p < 0.001$) and Room-Metaphor ($F(1, 597) = 4.58$, $p = 0.032$) feedback had significant main effects while the interaction effect between the two factors was not statistically significant ($F(1, 597) = 0.98$, $p = 0.323$). Basically, both approaches made people spend more time solving the problems, probably due to the increased interaction required by the additional features.

Second, we analyzed the time difference between Baseline Quiz and Quiz 1. Both Test-Case ($F(1, 597) = 49.67$, $p < 0.001$) and Room-Metaphor ($F(1, 597) = 4.40$, $p = 0.036$) feedback had significant main effects while the interaction effect between the two factors was not statistically significant ($F(1, 597) = 1.62$, $p = 0.204$). When the interface includes the Test-Case feedback, the time spent increases with an average of 49.8 seconds, and with the Room-Metaphor feedback, the time spent increases with an average of 14.3 seconds.

Third, we analyzed the time spent by problem type including both Baseline Quiz and Quiz 1. Problem type had a significant main effect ($F(3, 1396) = 22.13$, $p < 0.001$). After conducting a Tukey HSD Test for pair-wise comparisons of each problem type, we found that participants spent a longer time with Type 1 (150.4 sec) than with Type 2, Type 3, and Type 4 (for all pairs $p < 0.001$). Participants spent a longer time with Type 3 and Type 4 (102.8 and 100.8 sec, respectively) than with Type 2 ($p = 0.001$ and $p = 0.003$, respectively) and there was no difference between Type 3 and Type 4 ($p = 0.998$). Finally, participants spent the shortest

time with Type 2 with an average of 64.5 seconds.

5.3. Discussion

5.3.1. Data Collection through Crowdsourcing

The quality of the collected data was better than we had expected. The ratio of correct responses was comparable with the results of previous studies discussed in the Background section. Participants also spent a reasonable amount of time (average 50.0 to 171.1 seconds depending on conditions) solving the given questions. Only less than 5% of incorrect responses were difficult to understand, and most of the responses (over 95%) were either correct answers or incorrect answers with the reversal error. The dichotomous responses eased the burden of analyzing the collected data.

5.3.2. Effects of Different Interfaces

H1 – confirmed: The results of the statistical analysis show that the Test-Case approach is an effective way to reduce the reversal error. However, the same issues that arose when using the Test-Case approach in the interview study also appeared while conducting User Study 1. Some participants did not seem to fully understand why their responses did not match with their test cases. We assume they switched the variables to get the correct feedback. Here are some comments that participants made in the online survey after the experiment:

> *"As far as I was concerned, the interface forced me to write wrong answers to the question."*

> *"It would tell me that my numbers didn't match, but I didn't know another way to write them."*

Note that these types of comments were only found in the Test-Case condition, and *not* in the Both condition, even though the Both condition also employed the Test-Case approach. This could be supporting evidence that the Room-Metaphor approach helped research participants understand why there were mismatches.

H2 – confirmed: The Room-Metaphor approach also turned out to be effective in reducing the reversal errors. However, if one looks at the data closely, it appears that the Room-Metaphor approach is only effective in certain situations. As shown in Figure 6, when the Room-Metaphor condition was used without the Test-Case approach, the number of correct responses did not increase much between the Baseline Quiz and Quiz 1. Simply providing a visualization did not significantly perturb the participants' thinking. Considering that the time spent decreased while solving Quiz 1 in the Room-Metaphor condition compared to the Baseline Quiz, we assume that participants did not spend additional time to understand the visual feedback. This could be explained by constructivist theory, which promotes that proper engagement is essential for the visualization to be effective.

H3 – confirmed: There appears to be clear interaction effects between the two approaches. When both Test-Case and Room-Metaphor approaches were used together (the Both condition), greater improvement was observed. We believe that each of the two approaches served different purposes, so when they were used together, they complemented each other. First, the Room-Metaphor visualized the relationship of the mathematical equation. As discussed in Section 2.1.1, the reversal error could arise from a misunderstanding of the mathematical relationship between the variables of the equation. In order to prevent the consideration of variables as static labels, the Room-Metaphor forced participants to consider variables as representing different numbers of objects. This understanding is closer to an

expert's mental model of an equation, which should be an effective visualization based on epistemic fidelity theory. However, the visualization alone does not alert the participant to the incorrectness of the mental model. The effectiveness of the visualization increases when it is combined with the Test-Case approach. If there is a mismatch with the equation and sample values, the Test-Case approach indicates that there is something wrong by changing the size of the Room-Metaphor to be equivalent on both sides. After it triggers the participants to realize that there is a discrepancy, an understanding of *why* it is wrong appears to happen while engaging with the Room-Metaphor visualization. With feedback combined from the two approaches, learners can see how their sample values fit with their equation. It is more likely that one can understand the error while seeing how the visualization of the equation changes when interacting with the test-cases rather than purely simulating the mental model one has in her mind. During this process, we believe that if one truly understands the *why* component, the mental model can be reconstructed and properly updated.

Some results are certainly encouraging; while solving Type 3 and Type 4 problems, participants using both Test-Case and Room-Metaphor approaches were able to increase the number of correct responses by over 100% (from 12 to 24 with Type 3 and from 9 to 22 with Type 4). However, it is difficult to claim that these approaches completely resolve the reversal error issue: More than 50% of participants still responded incorrectly. This indicates how difficult it is to change the underlying mental models for even a simple mathematical word problem. The fact that this technique may make a difference for even a small number of learners, however, should be very encouraging for teachers and researchers in engineering and mathematics where this issue persists.

Participants using both the Test-Case and Room-Metaphor feedback spent more time solving the given problems. The longer time spent could be due to several factors, including just spending more time solving the problem, the inevitable interaction time that is required from the interface, or additional consideration after observing the feedback. Although it is not clear, we believe the Test-Case approach increased the time spent more than the Room-Metaphor approach as it required entering sample values and correcting the answer if the Test-Case approach informed the participant that the provided answer is incorrect.

5.3.3. Effects of Problem Types

We identified different patterns in the data for the different problem types. We see that the average correct response rate of Type 1 problems—problems that are essentially the same as the original student-professor problem—is 44%, which is within the range of results in the prior literature. Interestingly, the time spent was the longest with Type 1 problems. We believe that this is not due to the difficulty level of the problem, but a limitation of our experiment. We randomized the order of the four problems so they did not show up in the same order for Baseline Quiz and Quiz 1. However, this order was fixed for all of the participants, and a Type 1 problem was shown first. As it was the first trial, a longer time might have been required to learn the task and the interface.

For Type 2, the ratio of correct responses is relatively higher (average of 86% in all the conditions) than the other types of problems. This result also fits the expectations based on prior literature. We assume that this high ratio of correct responses is related to the objects that are used in the problem description, such as mountains and rivers. In the student-professor problem, the term "students" itself does not match with the variable S, which represents the *number of students*.

However, when we think about mountains and rivers, the objects' inherent heights and lengths may help the participants move away from thinking of the variables as labels. The relative easiness of this problem type can also be explained by the fact that participants spent the least time with Type 2 problems.

We see that little or no change was found in the number of correct answers for any of the feedback approaches for Type 1 or Type 2 problems in this round of data collection. Without interviews with the participants, it is difficult to fully explain why this occurred. However, further inspection of the data shows that a large majority of those participants who correctly answered the Type 1 or Type 2 questions in the Baseline Quiz also answered the corresponding Quiz 1 questions correctly (in fact, this was true for all but 4 participants). This suggests that these participants already understood these problem types at an anticipatory level (Tzur and Simon, 2004), and the different feedback prompts were unnecessary at these levels. These participants were able to solve a similar problem in the same way in both the Baseline Quiz and Quiz 1, and their understanding was neither interrupted nor aided by the prompts.

However, our feedback approaches did have significant effects on Type 3 and Type 4 problems, which are the problem types that traditionally present more difficulty. Participants spent more time with Type 3 and Type 4 problems. By looking at the data in more detail, we see that for Type 3, more than half of the participants who answered incorrectly on the Baseline question were able to answer correctly under the Test-Case or Both approach. Similarly, on Type 4 problems, more than half of those who initially answered incorrectly were successful when provided with the Both approach, although interestingly here, neither the Test-Case nor Room-Metaphor alone showed an improvement. For these more difficult

problems, we suggest that these participants had some understanding of how to solve these problems, but their understanding did not yet include an anticipation of the connection between the activity and effect of that activity (Tzur and Simon, 2004). Given a prompt, in this case in the form of the feedback given in the Both condition, they were able to engage successfully in the problem solving.

6. User Study 2

Though the results of Study 1 are quite encouraging, the effects of proposed approaches (especially, the Room-Metaphor approach) would be subject to the "next-day phenomenon" (Tzur and Simon, 2004). Some may also argue that the improved performances are simply the side effects of longer time spent on these additional interaction techniques. Thus, User Study 2 was designed to test whether the proposed techniques have lasting effects, specifically testing H4 and H5. If the suggested approaches truly alter the students' mental models, the learning should have longitudinal effects.

6.1. Methods

6.1.1. Participants

At the end of User Study 1, participants were asked to submit their email addresses if they were interested in participating in a follow-up study. Through the collected email addresses, we recruited a total of 63 participants (15 participants who participated in the Baseline condition in User Study 1, 15 in the Test-Case condition, 13 for the Room-Metaphor condition, and 21 for the Both condition). The participants' ages in User Study 2 ranged between 18 and 59, and the average age was 31. Each participant was compensated with $0.15 as a bonus payment.

6.1.2. Procedures

Following previous literature, we re-invited the participants after a period of at least two days, and a maximum of six days. All of the participants were again asked to solve the four types of problems in the Baseline condition. That is, the participants did not receive help from any of the interactive feedback approaches tested in this trial. Table 1 also shows the problems used in User Study 2, which we call "Quiz 2." Note that we did not reuse any problems used in the Baseline Quiz and Quiz 1 for Quiz 2.

6.1.3. Measures

Our primary goal for User Study 2 was to see if the improvement in User Study 1 was only shown within the context of using the interface, or if the participants understood the underlying concepts and could solve problems effectively without interactive aids. Although we invited all participants from User Study 1 to take part in User Study 2, in the data analysis we chose to only include participants who had improved between the Baseline Quiz and Quiz 1. If we included participants who did not show improvement previously (i.e., those who submitted all incorrect responses or all correct responses in User Study 1), the correct responses in Quiz 2 could have been possibly attributed to other factors that participants were exposed to between User Study 1 and User Study 2 or before User Study 1. Using the improvement metric used in User Study 1, we identified participants that improved on any of the four types of problems, and included that participant's data in User Study 2 for analysis. For this subset of participants, we count the number of correct responses in Quiz 2.

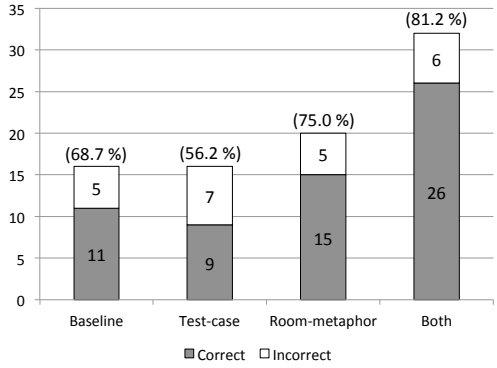


Figure 8: Number of correct and incorrect responses for Quiz 2 who had at least one improvement from User Study 1.

6.1.4. Results

The total number of participants who showed at least one improvement was 21 (4 participants who participated in the Baseline condition in User Study 1, 4 in Test-Case, 5 in Room-Metaphor, and 8 in Both), which is 1/3 of all the voluntary participants in User Study 2. All 21 of these participants answered four questions each for Quiz 2, giving a total of 84 responses to code.

Figure 8 shows the number of correct responses for Quiz 2 among participants who had at least one improvement in User Study 1. We employed logistic regression to see if there were any significant main effects and interaction effects. We found that neither the effects of the Test-Case approach (Wald Chi-Square = 0.2866, $p = 0.5924$) nor the interaction effects (Wald Chi-Square = 0.8054, $p = 0.3695$) had a significant impact on the correct responses. Only the effects of the Room-Metaphor approach (Wald Chi-Square = 3.2156, $p = 0.0729$) were marginally significant.

6.1.5. Discussion

H4 & H5 – unclear: Since the number of participants who volunteered for the follow-up study (User Study 2) was small and unbalanced among the four conditions, it is difficult to make any statistically sound conclusions from User Study 2. However, based on Figure 8 and the marginal significance of the Room-Metaphor approach, we might be able to say that the Room-Metaphor approach is potentially effective in helping students grasp a concrete conceptual understanding of *SP*-type problems.

Our conjecture is that the improvement shown in User Study 1 with only the Test-Case feedback may not be based on a complete understanding of the underlying concept. As also pointed out in the previous literature (Kaput et al., 1985; Tzur and Simon, 2004), these participants may have simply changed the answer to get the correct feedback message without knowing why their original answer was incorrect. This is in line with the next-day phenomenon where students could not independently recall the knowledge to solve the problem correctly. Participants who were exposed to the Room-Metaphor approach in User Study 1 had a success rate of 75%, while those exposed to the Both condition had a success rate at 81.25%. Due to the small sample size, we cannot claim any significant difference between the different approaches. However, as the both condition had a relatively higher proportion of correct responses, we believe that the Room-Metaphor approach has the potential to help the participants obtain the knowledge without being limited to the situation of having the feedback interface. In other words, the feedback interactions may have helped change their mental models by building an understanding of the underlying concepts.

7. Conclusions

We developed and evaluated an interactive visual component, called "POETIC," using two approaches (Test-Case and Room-Metaphor) to help students solve *SP*-type problems without committing the reversal error. The overarching research objective of this study is to test the effectiveness of the two approaches on reducing reversal errors. The results of two user studies show that some participants successfully avoided reversal errors using POETIC for specific types of problems. Although we failed to completely eradicate the reversal errors, we observed that proper feedback and well designed visualizations can invoke a learner to avoid the errors appropriately. Though this study only achieved marginal improvements on a finite set of problems, we believe that the results are still encouraging considering that errors with *SP*-type problems have been a long standing issue over the past three decades.

Through this study, we suggest that our two feedback approaches can help our participants address this issue and perturb their thinking in a way that makes them question their original ideas and change their way of thinking. Particularly, when both approaches are used together, we see that they have the potential to reduce the occurrence of reversal errors, and there is marginal evidence showing that they could create long-term change to mental models and mitigate the next-day phenomena. Slightly different effects of each approach also showed that different visualization and interaction approaches have different roles, which coincide with the two theories discussed: epistemic fidelity theory and constructivism.

We see great potential for our findings to provide useful insights for teachers of mathematics at all levels. The reversal error is one that has been witnessed with elementary algebra learners all the way up to university engineering students

and adult learners. The results of this study suggest that it may be productive for teachers to make use of a metaphor for variables as "containers" of objects to help break the static comparison strategy. They should also feel encouraged to push their students to anticipate specific values (test cases) for these variables that they believe should be true for a given relationship and reflect on how their anticipated values make sense to an equation that they generate to represent the problem. Making such anticipatory and reflective practices a part of students' regular mathematical practice may lead to better overall mental models of the structure of mathematical equations. Our work leads us to believe that the POETIC tool, a particular implementation of interactive visualization, may be one way of helping students build these practices.

In addition, we hope that the results of this study will also be informative to human-computer interaction and information visualization researchers, who are interested in impacting and changing pre-existing mental models. The different roles of interactive visualization identified in this study could help other researchers design their own interactive visualization techniques.

8. Future Work

In future studies, we will conduct an interview or focus group study with people who have used POETIC, so that we can observe uses and interactions with the different approaches in POETIC. In spite of various benefits of crowdsourcing-based user studies, a lack of direct interaction with research participants is one of the biggest limits to our understanding of how the learner interacts successfully with the computer tool. We hope that direct observation and interviews will alleviate this issue.

POETIC will be integrated into our larger, web-based, interactive mathematical optimization education tool, POET, which not only deals with errors in *SP*-type problems but also other errors that student run into while building complex mathematical optimization models out of word problems. Thus, other interactive and visualization approaches in addition to the two approaches discussed here are being developed and evaluated and hopefully will prove productive in further helping students with mathematical modeling practices.

It is our hope that these and other studies in the domain of mathematics education and mathematical optimization modeling will provide an interesting set of visual, interactive approaches and supporting empirical evidence. We strive to continue to deepen our understanding of how various interaction and visualization approaches can help change resilient mental models, to provide useful resources for both teachers and researchers, and to have broad impacts on various areas of education and human-computer interaction.

9. Acknowledgments

This material is based upon work supported by the Purdue Engineer of 2020 Seed Grant Program and the National Science Foundation (Grant No. 1044182). We also appreciate Dr. Yukiko Maeda for her thoughtful review and suggestions for our statistical analysis, and all of the human subjects who participated in this study.

References

Beck, K., 2003. Test-Driven Development: By Example. Pearson Education, Boston, MA.

Clement, J., 1982. Algebra word problem solutions: Thought processes underlying a common misconception. Journal for Research in Mathematics Education 13, 16–30.

Cohen, E., Kanim, S.E., 2005. Factors influencing the algebra "reversal error". American Journal of Physics 73, 1072–1078.

Fisher, K.J., Borchert, K., Bassok, M., 2010. Following the standard form: Effects of equation format on algebraic modeling. Memory & Cognition 39, 502–515.

Fisher, K.M., 1988. The students-and-professors problem revisited. Journal for Research in Mathematics Education 19, 260–262.

Hegarty, M., Mayer, R., Monk, C., 1995. Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. Journal of Educational Psychology 87, 18–32.

Hundhausen, C., Douglas, S., 1999. Toward effective algorithm visualization artifacts: designing for participation and communication in an undergraduate algorithms course. Ph.D. thesis. University of Oregon.

Hundhausen, C.D., Douglas, S.A., Stasko, J.T., 2002. A meta-study of algorithm visualization effectiveness. Journal of Visual Languages & Computing 13, 259–290.

Kaput, J.J., Clement, J., 1979. Letter to the editor. Journal of Children's Mathematics Behavior 2, 208.

Kaput, J.J., Sims-Knight, J.E., Clement, J., 1985. Behavioral objections: A response to wollman. Journal for Research in Mathematics Education 16, 56–63.

Kenney, R., Uhan, N., Yi, J.S., Kim, S., Gopaladesikan, M., Shamsul, A., Hundia, A., 2011. Understanding and overcoming difficulties with building mathematical models in engineering: Using visualization to aid in optimization courses, in: Research in Undergraduate Mathematics Education (RUME), Portland, OR.

Kittur, A., Chi, E.H., Suh, B., 2008. Crowdsourcing user studies with mechanical turk, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy. pp. 453–456.

Lawrence, A.W., 1993. Empirical studies of the value of algorithm animation in algorithm understanding. Ph.D. thesis. Georgia Institute of Technology.

Liu, Z., Stasko, J.T., 2010. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. IEEE Transactions on Visualization and Computer Graphics 16, 999–1008.

Lochhead, J., 1980. Faculty interpretations of simple algebraic statements: The professor's side of the equation. The Journal of Mathematical Behavior 3, 30–37.

MacGregor, M., Stacey, K., 1993. Cognitive models underlying students' formulation of simple linear equations. Journal for Research in Mathematics Education 24, 217–232.

Mulholland, P., 1998. A principled approach to the evaluation of SV: a case study in prolog, in: Brown, M., Domingue, J., Price, B., Stasko, J. (Eds.), Software Visualization: Programming as a Multimedia Experience. MIT Press, Cambridge, MA, pp. 439–452.

Palm, T., 2008. Impact of authenticity on sense making in word problem solving. Educational Studies in Mathematics 67, 37–58.

Philipp, R.A., 1992. A study of algebraic variables: beyond the student-professor problem. Journal of Mathematical Behavior 11, 161–176.

Resnick, L., 1989. Knowing, learning, and instruction: Essays in honor of Robert Glaser. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Sims-Knight, J.E., Kaput, J.J., 1983. Exploring difficulties in transforming between natural language and image based representations and abstract symbol systems of mathematics. The Acquisition of Symbolic Skills , 561–570.

Stacey, K., McGregor, M., 1993. Origins of students' errors in writing equations, in: Baturo, A., Cooper, T. (Eds.), New Directions in Algebra Education. Queensland University of Technology, Brisbane, Australia, pp. 205–212.

Stasko, J., Badre, A., Lewis, C., 1993. Do algorithm animations assist learning?: An empirical study and analysis, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands. pp. 61–66.

Stokes, M.E., Davis, C.S., Koch, G.G., 2000. Categorical data analysis using the SAS system. SAS Institute Inc., Cary, NC.

Stylianou, D.A., 2002. On the interaction of visualization and analysis: the negotiation of a visual representation in expert problem solving. The Journal of Mathematical Behavior 21, 303–317.

Tung, S., Chang, C., Wong, W., Jehng, J., 2001. Visual representations for recursion. International Journal of Human-Computer Studies 54, 285–300.

Tzur, R., Simon, M., 2004. Distinguishing two stages of mathematics conceptual learning. International Journal of Science and Mathematics Education 2, 287–304.

Vredenburg, K., Mao, J.Y., Smith, P.W., Carey, T., 2002. A survey of user-centered design practice, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Minneapolis, MN. pp. 471–478.

Waisel, L.B., Wallace, W.A., Willemain, T.R., 2008. Visualization and model formulation: an analysis of the sketches of expert modellers. Journal of the Operational Research Society 59, 353–361.

Weinberg, A., 2009. Students' mental models for comparison word problems, in: Swars, S.L., Stinson, D.W., Lemons-Smith, S. (Eds.), Proceedings of the 31st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Atlanta, GA. pp. 709–717.

Wollman, W., 1983. Determining the sources of error in a translation from sentence to equation. Journal for Research in Mathematics Education 14, 169–181.

Yazdani, M., 2008. The limitations of direct sentence translation in algebraic modeling of word problems. Journal of Mathematical Sciences & Mathematics Education 3, 56–61.